NSW GOVERNMENT Health

Vol. 23(1–2) • January–February 2012
ISSN 1034 7674

NSW PUBLIC HEALTH BULLETIN

Health Statistics NSW: getting the right balance between privacy and small numbers in a web-based reporting system

***James P. Scandol*[A,B] *and Helen A. Moore*[A]**

[A]*Centre for Epidemiology and Research, NSW Ministry of Health*
[B]*Corresponding author. Email: james.scandol@doh.health.nsw.gov.au*

Abstract: Health Statistics NSW is a new web-based application developed by the Centre for Epidemiology and Research at the NSW Ministry of Health. The application is designed to be an efficient vehicle for the timely delivery of health statistics to a diverse audience including the general public, health planners, researchers, students and policy analysts. The development and implementation of this web application required the consideration of a series of competing demands such as: the public interest in providing health data while maintaining the privacy interests of the individuals whose health is being reported; reporting data at spatial scales of relevance to health planners while maintaining the statistical integrity of any inferences drawn; the use of hardware and software systems which are publicly accessible, scalable and robust, while ensuring high levels of security. These three competing demands and the relationships between them are discussed in the context of Health Statistics NSW.

Health Statistics NSW (HSNSW, www.healthstats.nsw.gov.au) is a new web-based reporting application developed by the Centre for Epidemiology and Research, New South Wales (NSW) Ministry of Health. This application was developed to replace the electronic version of the report *Health of the people of NSW—Report of the Chief Health Officer* (known as the eCHO report) which has been published as a web-based report since 2000 and as a printed report since 1996. The main incentives for converting the eCHO report, which consisted of a series of around 1000 static web pages, into an interactive web-based application were:

- to assist users find specific information through improved search and navigation functions
- to assist users download data and reports through improved download and report compilation functions
- to expand the content, update data and maintain the report through the use of modern business intelligence software, infrastructure and technology.

Many Australian (Australian Institute of Health and Welfare, aihw.gov.au, Public Health Information Development Unit, www.publichealth.gov.au and the Australian Bureau of Statistics, www.abs.gov.au) and overseas health agencies (US Centers for Disease Control and Prevention www.cdc.gov/DataStatistics/ and the World Health Organization, www.who.int/research/en/) now incorporate some type of web-based data query system into their main agency website.[1] The rapid development of web-based technologies in the past 20 years has seen such sites evolve from simple tables of data to hosting complex web-based applications that allow significant user interaction including the production of dynamically generated graphs and maps. For example, new map-based reports using InstantAtlas™ (www.instantatlas.com) are used by the Victorian Department of Health (www.health.vic.gov.au/healthstatus/atlas/) and to report the Australian Early Development Index (maps.aedi.org.au/IA/2011/region/105/atlas.html).

CONTENTS – See back page
www.publish.csiro.au/journals/phb

10.1071/NB11047

This trend of improving the technologies that support these types of websites is likely to continue into the foreseeable future with significant investment into the Australian internet infrastructure by both private and public institutions (for example the Australian National Broadband Network, www.nbnco.com.au).

This article describes the three competing demands that were considered when designing the HSNSW application, configuring the content and then deploying a system which was suitable for release to the public. These demands were: consideration of the public versus the private interest when reporting health statistics; recognition of statistical signals versus noise when reporting at small spatial scales and on rare conditions; and building data systems that are secure but still highly accessible. An expanded consideration of these issues was presented in the report *Privacy issues and the reporting of small numbers*.[2]

1. Public versus private interest

The implementation of evidence-based policies and planning for health services requires the collection and management of data. Development of performance indicators from these data supports our understanding of whether particular policies and programs are achieving their goals; there is also significant public interest in performance reporting. Further, all levels of government require the use of timely data for the planning of clinical and public health services.[3,4] In the field of health however, we must be particularly cognisant that these data are collected from individuals and that there are significant legal and ethical reasons why the privacy of these individuals cannot be compromised. Consequently the overarching challenge when presenting health statistics is to develop robust reporting strategies that ensure that both private and public interests are met.

Within the NSW health system, the key legislative instruments to protect the privacy of citizens are the *Privacy and Personal Information Protection Act 1998* (NSW) which regulates personal information in the public sector; and the *Health Records and Information Privacy Act 2002* (NSW) which regulates personal health information. The *Health Records and Information Privacy Act 2002* is supported by detailed statutory guidelines which cover particular applications of health data. All relevant laws and policies are explained in the *NSW Health Privacy Manual*.[5]

A critical aspect of privacy legislation when using health data is the de-identification of the data. De-identification in the context of public reporting must be interpreted more broadly than simply removing names and addresses from records because it is about the potential of re-identifying an individual from the final publication of that data. The steps required for effective de-identification in this context are not necessarily simple and require consideration of the condition being reported and the population from which observations are drawn. Much emphasis in public reporting is usually placed on the number of people reported and simple threshold rules are defined. However, expert groups such as the Statistical Information Management Committee[6] argue that the size of the underlying population (which may be defined as the population in one geographic area or a sub-group such as the Aboriginal population) becomes more important when the probability of re-identification is considered.

Consequently when HSNSW was configured, standard rules were used to guard against re-identification (such as designing tables to minimise the number of cells with denominators less than 1000 people and individual counts less than five people). Such steps are crucial to ensure that private interests are not compromised to achieve the public interests associated with statistical reporting.

2. Data signals versus noise

HSNSW includes partial functionality to drill down into increasing levels of data granularity (the fineness with which data fields are subdivided). For example, when looking at hospitalisation admissions, it is possible to examine the pattern of these admissions across Local Health Districts, or to develop a time series of admissions for a particular Local Health District. This functionality was included because many potential users of the system requested straightforward access to data about their Local Health District. This approach works well but it very rapidly becomes apparent that there are limitations to how far you can drill down into the data before the numbers of individuals being reported become too small to meet two important criteria: privacy and statistical interpretability. Firstly, as noted above, there are privacy issues that cannot be compromised. Secondly, as Steel and co-workers discuss,[7] small numbers are subject to much larger relative variation over time or between groups, which makes any inferences drawn from these numbers less reliable. It is important to recall that the reason these statistics are being reported in the first place is the public interest associated with evidence-based policy and decision making. If the inferential value of these data is degraded, then the justification for their publication can become compromised.

Perhaps the most transparent approach to this issue is to estimate the variability of any published statistics (which may require additional statistical assumptions). If the relative variability exceeds some particular threshold, then such statistics should not be published. Where practical, this approach was used in HSNSW, but in other cases, simple techniques such as averaging or more sophisticated statistical methods such as Bayesian smoothing[8] were used to ensure that patterns, and not observations from individuals, were being presented. Judgments based upon

threshold rules of sample size and relative variability, which were checked by subject matter experts, were used when configuring indicators for HSNSW to ensure that there were adequate data for any meaningful statistical inferences. Options to drill down into these data were not provided if such subsets of the data were subject to excessive variability (e.g. for mesothelioma deaths, which are very rare events). If analysts, planners or policy developers require such data then there are alternative options for accessing this information within secure, non-public environments (such as SAPHaRI).[9]

Note that technologies designed for data drill-down such as data-cubes (which can be thought of as a multi-dimensional extension to a spreadsheet table) have primarily been developed for commercial applications. For example, a sales manager may want to see patterns of sales across the nation, but he or she may also want to know what type of widget a particular salesperson sold yesterday. Such datasets and applications are not associated with the privacy and inferential issues that are so important to the health data being presented with HSNSW. Although HSNSW does use data-cubes to efficiently access large volumes of data, decisions about the level of data granularity available to the public are made well before the data are transformed into cubes.

3. Data security versus accessibility

The underlying technical architecture of HSNSW is complex and the details are beyond the scope of this article. There are, however, two major components of the system: indicator calculation; and reporting and analytics. The algorithms used for indicator calculation process unit-level or semi-aggregated data into defined health indicators on secure internal workstations. These processing steps are implemented using existing and well-tested processes within the Centre for Epidemiology and Research, NSW Ministry of Health. Any data with potential privacy issues are therefore subject to strict security protocols within the Ministry. None of these privacy-sensitive data are stored on publicly-facing servers.

In contrast, the web-based reporting and analytics solution imports the text output from the indicator calculation steps described above, builds data-cubes, handles user interaction and renders tables, charts, maps and portable document format (pdf) reports, spreadsheets and images. These functions are completed on other servers which are publicly facing and are isolated from any servers which contain privacy-sensitive data.

This system thus enables the delivery of population health indicators to the public without any privacy-sensitive data being stored on public web servers. This design required some duplication and inefficiencies (from a systems-design perspective), but these are justified to meet the dual objectives of secure health data and accessible health indicators.

Conclusion

Web-applications such as HSNSW are complex systems which require consideration of a diverse range of issues in their design, implementation and configuration. Many of these issues require a trade-off between users' wishes and the responsibilities of the data reporting specialists. For example, people want access to data, but cannot be given access to all data because of very justifiable privacy issues. People need access to information about their area, but should not be provided with information that is not suitable for drawing valid statistical inferences. Computer systems need to be deployed that provide public access to the data, but these systems must be designed in a manner that cannot increase risks to personal privacy.

The authors contend that HSNSW strikes the right balance with these inter-related competing demands for the benefit of publishing a diverse range of population health indicators using a new web-based data query system. These issues are discussed in more detail in the report *Privacy issues and the reporting of small numbers*[2] that was prepared in conjunction with the initial release of HSNSW. Readers may also find the article by Lawlor and Stone[10] of interest as these authors provided an overview of tensions between data protection and informing public health.

Acknowledgments

The authors wish to acknowledge the following staff from the Ministry of Health who worked on the development of the HSNSW application: Lina Persson, Nicole Mealing, Mark Cerny, Hanna Noworytko, Christian Allen, Sarah Thackway, Kerry Chant, Dejan Mirkovic, Greg Thompson, Sachida Ghimire and Dudley Collinson. The authors are also grateful to the two reviewers whose comments improved the manuscript.

References

1. Rudolph BA, Shah GH, Love D. Small numbers, disclosure risk, security, and reliability issues in web-based data query systems. *J Public Health Manag Pract* 2006; 12: 176–83.
2. Centre for Epidemiology and Research. Health Statistics NSW: Privacy issues and the reporting of small numbers. Sydney: NSW Department of Health; 2011. p. 18.
3. COAG Reform Council National Healthcare Agreement. Performance report for 2009–10. Sydney: COAG Reform Council; 2011.
4. Government NSW. NSW 2021: A plan to make NSW number one. Sydney: NSW Government; 2011.
5. NSW Health. NSW Health Privacy Manual, Version 2, PD2005_593. North Sydney: NSW Ministry of Health; 2005. p. 75.
6. Statistical Information Management Committee (SIMC). Guidelines for the use and disclosure of health data for

statistical purposes. Canberra: Australian Institute of Health and Welfare; 2007. p. 16.

7. Steel D, Green J, Brown L. Best practice in small area analysis and reporting – literature review and guidelines. Centre for Health Service Development, University of Wollongong; 2003. p. 108.

8. Lawson AB, Browne WJ, Rodeiro CL. Disease Mapping with WinBUGS and MLwiN. Chichester: John Wiley & Sons; 2003. p. 292.

9. NSW Health. Secure analytics for population health research and intelligence (SAPHaRI). Sydney: NSW Health; 2011.

10. Lawlor DA, Stone T. Public health and data protection: an inevitable collision or potential for a meeting of minds? *Int J Epidemiol* 2001; 30: 1221–5. doi:10.1093/ije/30.6.1221